Lexical Retuning Targets Features

Karthik Durvasula and Scott Nelson
Michigan State University

1 Introduction

The speech perception mechanism must be equipped with a way to categorize ambiguous, unclear, or novel pronunciations of segments a listener hears. A series of experiments beginning with Norris et al. (2003) have used lexical retuning, wherein ambiguous segments presented in a lexical environment lead to a shift in listeners' categorical boundaries, to show one way in which the speech perception mechanism handles less than perfect input. Our paper will expand upon the lexical retuning literature by showing that lexical retuning targets sub-segmental (potentially, featural) representations.

In their landmark study, Norris et al. (2003) presented listeners with an auditory lexical decision task that included certain training items containing an ambiguous segment that was a blend of [f] and [s]. This ambiguous segment, [?$_{fs}$], was then used to create two different versions of their exposure list. The first list replaced all of the natural [f] occurrences with [?$_{fs}$] and kept all of the [s] occurrences normal, while the second list did the opposite, i.e., [f] occurrences remained normal and words containing [s] were replaced with [?$_{fs}$]. All of these training words were controlled so as to not form a minimal pair if replaced with the other segment of the pair [f s]. This allowed for a specific lexical target for the listener to analyze the ambiguous segment in. The remainder of the words presented in the lexical decision task were a mix of real and nonce Dutch words that contained no instances of [f s v z].

After being exposed to the ambiguous sound in the lexical decision task, participants were given a forced choice phonetic categorization task of sounds from a continuum containing varying proportions of [f] and [s] in coda position following an [ɛ] vowel. Participants who heard [?$_{fs}$] in the words normally containing [f] responded with "f" more often, therefore shifting their continuum to be more f-like. Likewise, participants who heard [?$_{fs}$] in the words normally containing [s] responded with "s" more often, shifting their continuum to be more s-like. Additionally, there was a third group that heard [?$_{fs}$] only in the context of the nonce words. Participants who were given this condition responded between the [f]-group and [s]-group and therefore it was claimed that retuning did not occur and can only be facilitated by the placement of an ambiguous segment within a lexical item.

While these findings show that systematic retuning does occur, there has been a continual push since then to discover more about where in the grammar the retuning occurs and at what level. In a direct follow up to Norris et al. (2003), McQueen et al. (2006) kept the auditory lexical decision task the same, but replaced the phonetic categorization with a cross-modal priming task containing test items that formed minimal-pairs with [f]/[s]. All of the test items in the priming task were new to the participants. Listeners who were trained with [?$_{fs}$] in [f]-words were quicker to respond with "f" and those who heard it [?$_{fs}$] in [s]-words were quicker to respond with "s". Because all of these test items were novel, there is no (direct) way there could have been any acoustic trace of the ambiguous segment within the items and therefore it was argued that the locus of retuning must be pre-lexical (i.e., affect some sort of abstract representation).

Additional work around the time showed that the retuning effect was dependent on similarity between the training speaker and the test speaker (Eisner & McQueen, 2005; Kraljic & Samuel, 2005). Eisner & McQueen (2005) combined pairs of vowels and fricatives in different combinations from three different talkers to use in the phonetic categorization section of their experiments: two female and one male. They found that retuning only occurred when the fricatives used in testing came from the same speaker as the one used in training.

*Authors' names appear in alphabetical order. We would like to thank John Kingston, Yen-Hwei Lin, Amanda Rysling, audiences at GLEEFUL 2017 and AMP 2017, and the Phonology-Phonetics group at Michigan State University for insightful comments, questions and discussion on various aspects of this work. We would also like to thank Russ Werner for technical help and Julia Andary for help with creating stimuli.

Proceedings of *AMP 2017*

It did not matter if the vowel used in the test portion was from a different speaker than used in the training, their results depended exclusively on the fricative portion matching. The retuning effect was only found to occur when the fricatives were specifically from the same speaker. Kraljic & Samuel (2005) ran similar experiments but found that retuning could generalize across speakers if the spectral energy was distributed similarly in fricatives from different speakers. While the results from Eisner & McQueen (2005) suggest that the retuning effect is speaker-specific, the results from Kraljic & Samuel (2005) indicate that it is the phonetic properties that are being targeted by retuning.

Further results from Kraljic & Samuel (2005), as well as Eisner & McQueen (2006), show that the retuning effect does not diminish over time and only resets after the participant is presented with "normal" tokens of the previously ambiguous segment. These results are challenged by Van Linden & Vroomen (2007) who do a comparison of lexical retuning with visual recalibration of auditory speech. Visual recalibration uses a McGurk effect (McGurk & MacDonald, 1976) to shift the categorical boundaries producing similar results to lexical retuning (Bertelson et al., 2003). The results from Van Linden & Vroomen (2007) show that the effects from lexical retuning and visual recalibration are (at least) similar, if not the same. Additionally, they found that both effects diminished over time. Of note is the fact that they used stop consonants in their study while both Kraljic & Samuel (2005) and Eisner & McQueen (2006) used fricative consonants.

Variation in retuning between fricative-based and stop-based has already been noted in the literature. Kraljic & Samuel (2006) performed a lexical retuning study using stop consonants. They tested two things: 1) does the retuning effect generalize to new speakers since the cue for stops was less variable and 2) can listeners be trained using voiced stops and transfer the effects to a continuum of voiceless stops. A retuning effect was found for both the voiced and voiceless continua. Kraljic & Samuel (2006) took these findings to suggest that the speech perception mechanism behaved differently depending on the type of input it receives; particularly, when presented with stop consonants, the system can generalize down to the feature level due to the typical invariance of the acoustic cue, and therefore apply across the board. Though they did not directly test if retuning generalizes from a voiceless continuum to a voiced continuum for fricatives, they suggested that because the spectral cues used in fricative perception is more variant across individuals, the speech perception mechanism might be more resistant to broad generalizations when presented with ambiguous input of this type.

In the current study, we probe whether or not perceptual retuning targets features even in the case of fricative segments. Specifically, we probe if retuning in a voiceless [f]~[s] continuum generalizes to a voiced [v]~[z] continuum. This according to us is likely given that there has been a lot of work that shows that perceptions taps into features via perceptual confusion (Miller & Nicely, 1955) and selective adaptation (Eimas & Corbit, 1973; Eimas et al., 1973). Some have even argued that patterns in which categorical boundaries are perceived is consistent with feature categories as opposed to segmental categories (Chládková et al., 2015). If the retuning effect does indeed target features even for fricatives, then we should see that retuning in a voiceless [f]~[s] continuum indeed generalizes to a voiced [v]~[z] continuum without explicit training, i.e., even though the training materials are limited to an ambiguous segment [$?_{fs}$] between an [f] and an [s], and participants never hear any instance of [v] or [z] (outside of the phonetic categorization tasks), participants should generalize the retuning to the [v]~[z] continuum. We show that this prediction is borne out in what follows.

2 Methods

The methodology used in our study largely follows that of Norris et al. (2003) - a lexical decision task (LDT) is followed by phonetic categorization. This has been the standard throughout much of the lexical retuning literature (Eisner & McQueen, 2005; Kraljic & Samuel, 2005; Jesse & McQueen, 2011). The primary difference in our design is the addition of a pre-exposure (i.e., pre-LDT) phonetic categorization. This is not an entirely new design as Eisner & McQueen (2006) used a pre-training phonetic categorization task to show that participants' categorization functions were similar before receiving training. While they did not use their data to compare performance within groups, the addition of a pre-exposure phonetic categorization task also allows for within-subject "before" and "after" comparisons which we use in our analysis. The remainder of this section will briefly describe our methodology.

2.1 *Participants* 71 undergraduate students (Mean Age = 20.1 years; 51 female, 2 unreported gender) at Michigan State University took part in the experiment for extra credit. All participants identified as native English speakers and did not report any hearing problems.

2.2 *Design* The experiment consisted of three blocks. This included a pre-exposure phonetic categorization task, an auditory lexical decision task (LDT) where participants were exposed to the ambiguous segment, and a post-exposure phonetic categorization task. The pre-exposure and post-exposure categorization tasks were identical for each participant. Participants were assigned either the voiceless [f]~[s] or voiced [v]~[z] continua for their categorization tasks.

2.3 *Materials*

2.3.1 *Lexical Decision Task* A list of 75 English words and 75 phonotactically licit English nonce words was created to be used in the lexical decision task. Thirty-four of the real English words were used as training items designed to facilitate the lexical retuning. These were monosyllabic words containing either the sound [f] or [s]. Crucially they did not form a minimal pair with the opposing segment (*fool, cliff, soon, less*). Of the 34 words, 17 contained [f] and 17 contained [s]. Nine of the 17 words had the crucial segment appearing in the onset position while in the remaining eight words it appeared in the coda position. Test words were controlled for frequency using the subtlexus corpus (Brysbaert & New, 2009). While the [f]-words were less frequent (12.85/million) than the [s]-words (20.77/million), no statistically significant difference between the log/raw frequencies was found [$t(28.3)=0.48$, $p = 0.64$].

The remaining 114 words were used as fillers. The 41 normal English words and 75 nonce words varied in syllabic length from one to four syllables. None of the filler words contained any instances of [f s v z]. While it is standard to exclude these segments within the retuning literature, it was especially important for us to exclude any instances of the voiced fricatives. Our prediction crucially relies on participants in the "voiced" condition never being directly exposed to any versions of the voiced segments. If the retuning effect targets features then it should generalize from the ambiguous voiceless segments onto the matching voiced segments.

All 150 words used in the lexical decision task were spoken by a female native American English speaker from Michigan. The words were read aloud into a microphone in a quiet room and recorded directly into Praat (Boersma & Weenink, 2016) at a sampling rate of 44.1 kHz. For the nonce words, the speaker was presented the word using English orthography and asked to produce it in a naturalistic fashion. Each word and non-word was recorded twice and the impressionistically better of the two versions was subsequently chosen for use in the experiment. These chosen versions of both words and non-words were then isolated and saved as individual sound files. The 17 (crucial) real words containing [f] were further manipulated to replace the natural [f] with an ambiguous segment [$?_{fs}$]. This was done by identifying the location of the frication energy within the original word, removing it at the zero-crossing points closest to their surrounding environments, and replacing it with the ambiguous sound. Praat was used to automate this process and the resulting stimuli were manually checked to ensure they were natural sounding. Details on the creation of [$?_{fs}$] will be discussed below.

2.3.2 *Phonetic Categorization* In addition to recording the words for the lexical decision task, the speaker recorded tokens of [fi], [si], [vi], and [zi]. Two separate 41-step continua were created by blending the fricative section of the two sounds together using a Praat script. The voiceless continuum was created by isolating the fricative portion of [fi] and [si], matching them in length (165 ms) and intensity (65 dB), and blending them together by altering the amplitudes of each segment in equal steps along the continuum. Step 1 was therefore 100% [f] and 0% [s], while step 41 was 0% [f] and 100% [s]. Intermediate steps varied by 2.5% in opposite directions for both segments. All of the resulting sounds were then spliced onto the [i] vowel from the original recording of [fi] to create a full continuum. The voiced continuum was created the same way using [vi] and [zi] and only varied in the matched length (143 ms) of the blended fricatives.

A subset of each 41-step continuum was chosen to be used in the phonetic categorization tasks. Each subset was made up of steps 1 (100% labiodental) and 41 (100% alveolar) as well as 12 evenly spaced steps throughout the middle of the continuum. An ambiguous zone was determined by the authors for both the [f]~[s] and [v]~[z] continua. These zones were not the same for the two continua and therefore different steps

were chosen for each continuum. For the voiceless continuum, every second step between 7-29 was chosen. For the voiced continuum, the ambiguous zone was found to be later in the continuum and therefore every second step between 13-35 was used.

Both continua were then used in a pre-experiment designed to find the most ambiguous segment. 13 American English speakers (Mean Age = 20.9 years; 8 female, 1 unreported gender) separate from those used in the main experiment participated for extra credit. None of them reported any hearing problems. Each participant was tasked with categorizing both continua. The categorization of the voiceless continuum always preceded the categorization of the voiced continuum. For both continua, all 14 steps were played four times in random running order. Participants were tested using PscyhoPy (Peirce, 2007). They were given a forced choice task and asked if the sound they heard contained and "f" or "s" (for the voiceless continuum) and "v" or "z" (for the voiced continuum). After hearing a sound, they were instructed to use a computer mouse to click which sound they heard. The mean response for each step was calculated to create the two continua and the step at which the responses were closest to chance (50%) was taken to be the most ambiguous segment.

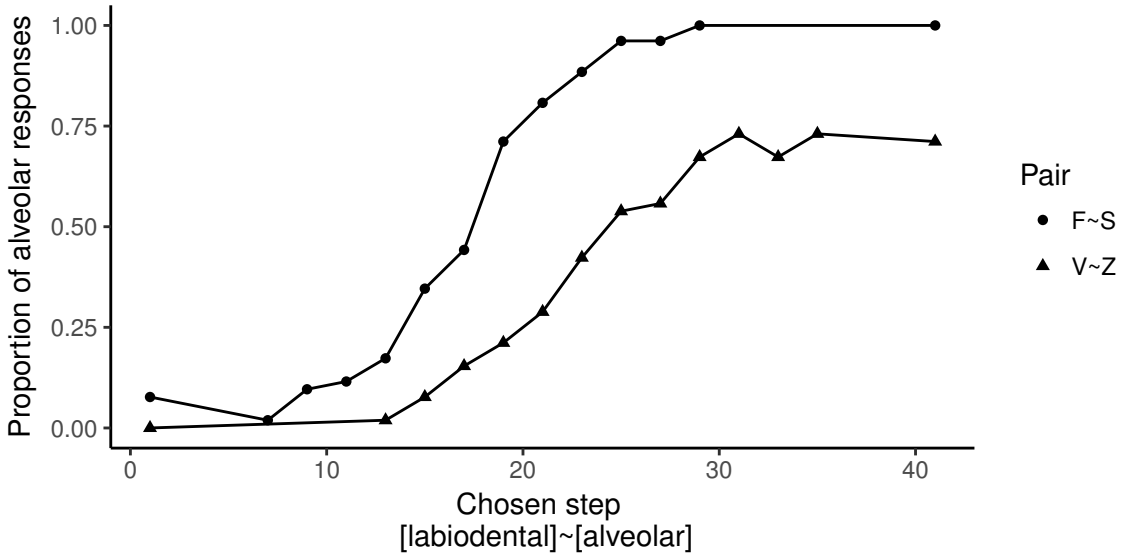

Figure 1: Categorization results for both continua from pre-experiment

The results of the pre-experiment show the expected sigmoidal categorical identifications for consonant segments (Figure 1). The results from the [f]~[s] continuum show that the response rate of 50% lies somewhere between steps 17 and 19. Therefore, we designated step 18 as [?$_{fs}$]. This is what was used to create the ambiguous versions of the f-words in the lexical decision task. We did not create an ambiguous voiced segment ([?$_{vz}$]) as none of our stimuli required this, but were still interested in what the baseline categorical function looked like for the voiced stimuli. Unlike the voiceless counterpart, the voiced-segment continuum never reached ceiling on the alveolar end of the spectrum. This shows an inherent bias towards the labiodental response that is strengthened by the fact that the 50% response rate does not occur until around step 24. It took listeners longer to cross over the categorical boundary and when they did they appear to have never accepted any of the "z"-like sounds completely.

2.4 *Procedure* Up to six participants were tested simultaneously in a quiet room. All stimuli were presented over headphones at an individual computer for each participant using the experimental software Psychopy (Peirce, 2007). Prior to the experiment, participants were verbally instructed that they would be doing three tasks on the computer using various response mechanisms - the phonetic categorization (first and third tasks) required the clicking of the mouse while the lexical decision task (second task) would require them to use the keyboard to give a yes/no response. Participants were also verbally instructed to answer as quickly and accurately as possible and to remain quiet until everyone in the room had completed all three tasks. All these instructions were also visually presented to them on the screen before each task.

The phonetic categorization tasks in the main experiment were similar to the ones used to find the ambiguous point. The crucial difference was that each participant in the main experiment only heard one of the continua. They were randomly assigned either to the voiceless continuum condition or voiced continuum condition. The same continuum was then used for both the pre-exposure and post-exposure phonetic categorization task. Each participant heard all 14 steps in their assigned continuum four times each in randomized order. They were given a forced choice task and instructed to use a computer mouse to click which sound they heard.

Upon completion of the pre-exposure categorization task, participants were instructed through Psychopy that they would now be hearing words one at a time and had to decide if what they heard was a real word of English. Participants were instructed to use the computer keyboard to respond "yes" or "no" to the question, "Is this an English word?" An 'a' response corresponded to "no" while an 's' response corresponded to "yes". If they did not respond within 3.5 seconds of the onset of the sound, no response was stored and the next sound was played. All words were presented in a randomized order.

After finishing the lexical decision task, participants repeated the same phonetic categorization task that they did prior to the lexical decision task. Everything in the second phonetic categorization task was identical to the original except that a new randomized order was generated.

3 Results & Discussion

Before looking at the main results related to the identification task, it is important to confirm that the participants did indeed accept the crucial test words in the LDT. This is important because retuning is contingent on the [?$_{fs}$] token appearing in real words, and therefore is contingent on participants accepting the words with the [?$_{fs}$] (replacing [f] in f-words) and the s-words as real words. Two participants were removed for failing to accurately judge the test items at a greater than 50% rate (threshold follows Norris et al. (2003)). The remaining participants did indeed have a reasonably high percentage of correct responses for the critical test words (f-words with [?$_{fs}$] = 83%; s-words = 87%). As a consequence, we should expect re-tuning to be possible.

A visual inspection of the proportions of alveolar responses in the pre-exposure ("Before") and post-exposure ("After") categorization tasks revealed that there was an observable decrease for the voiceless continuum (Figure 2), and for the voiced continuum (Figure 3).

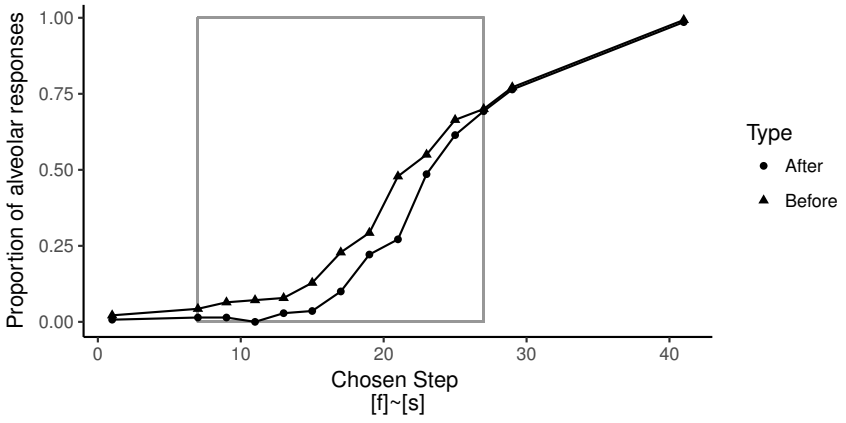

Figure 2: Before/After results for [f]~[s] continuum when [?$_{fs}$] appears in "f"-words

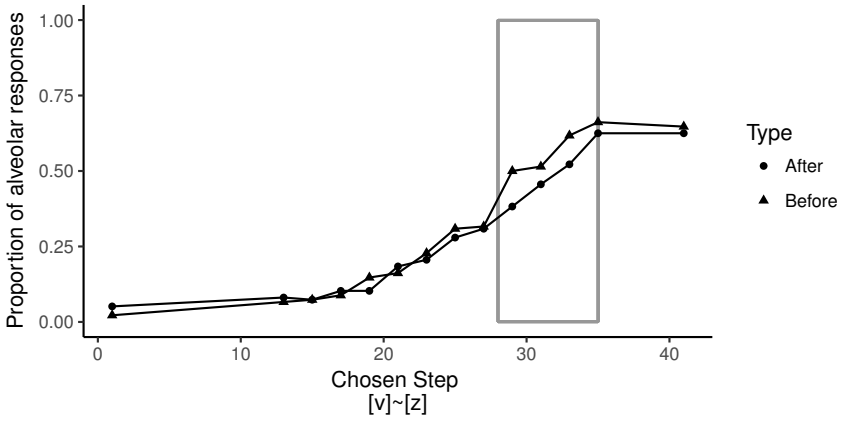

Figure 3: Before/After results for [v]~[z] continuum when [$?_{fs}$] appears in "f"-words

In order to confirm the observations made by visual inspection, we followed up the visual inspection with statistical analysis in R (R Development Core Team, 2014). We conducted paired one-tailed Mann Whitney U tests, with the proportion of alveolar responses as the dependent variable, and the timing of the identification experiment (Before vs. After) as the independent variable. These non-parametric tests were conducted as the dependent variable was a proportion, and hence the assumption of normality made by t-tests was violated.

We first looked at the results related to the [f]~[s] continuum. There was a statistically significant decrease in the overall proportion of alveolar responses for the voiceless ([f]~[s]) continuum in the post-exposure categorization task compared to the pre-exposure categorization task [$\text{MeanDiff}_{\text{Before-After}}$ = -6.1%, W = 104.5, p <0.001]. We then focussed on the region of the continuum near the 50% identification point in the pre-exposure categorization task, where the effect looked strongest (steps 7-27). For this region too, there was a statistically significant decrease in the proportion of alveolar in the post-exposure categorization [$\text{MeanDiff}_{\text{Before-After}}$ = -8.7%, W = 58.5, p <0.001]. The result is a replication prior results in showing that there is indeed an effect of lexical tuning on the [f]~[s] continuum due to the replacement of the [f] in the f-words with [$?_{fs}$].

We then focussed on the results related to the [v]~[z] continuum. For this continuum, there was no statistically significant decrease in the overall proportion of alveolar responses in the post-exposure categorization task [$\text{MeanDiff}_{\text{Before-After}}$ = -2.5%, W = 194, p =0.15]. However, when we focussed on the region of the continuum near the 50% identification point in the pre-exposure categorization task, where the effect was most pronounced during the visual inspection (steps 28-35), there was indeed a statistically significant decrease in the proportion of alveolar responses in the post-exposure categorization task [$\text{MeanDiff}_{\text{Before-After}}$ = -9.1%, W = 98, p =0.025]. The result suggest that replacement of the [f] in the f-words with [$?_{fs}$] affects the identification function of the [v]~[z] continuum.

The results here show that the replacement of [f] in f-words with [$?_{fs}$] affects both the [f]~[s] and [v]~[z] continuum (albeit, impressionistically, it affects the latter to a lesser degree). This in turn suggests that the lexical retuning is targeting sub-segmental features, potentially the place features [alveolar/labiodental].

4 Conclusion

In this article, we have shown that lexical retuning of the [f]~[s] continuum generalizes to the [v]~[z] continuum without any explicit training for the latter. This suggests that lexical retuning (also) targets a sub-segmental, potentially featural, representation with fricatives. Before concluding, it is important to point out that while we can be sure that the target is indeed sub-segmental, the experiment does not allow us to tease apart *phonological* features from *auditory* features. And we leave this for future work.

References

Bertelson, Paul, Jean Vroomen & Béatrice De Gelder (2003). Visual recalibration of auditory speech identification: a mcgurk aftereffect. *Psychological Science* 14:6, 592–597.

Boersma, Paul & David Weenink (2016). *Praat: doing phonetics by computer [Computer program]*. Version 6.0.19, retrieved 13 June 2016 from http://www.praat.org/.

Brysbaert, Marc & Boris New (2009). Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods* 41:4, 977–990.

Chládková, K., P. Boersma & T. Benders (2015). The perceptual basis of the feature vowel height. *Proceedings of XVIII ICPhS 2015 (article 711). Glasgow* .

Eimas, Peter D. & John D. Corbit (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology* 4:1, 99 – 109, URL http://www.sciencedirect.com/science/article/pii/0010028573900066.

Eimas, Peter D., William E. Cooper & John D. Corbit (1973). Some properties of linguistic feature detectors. *Perception & Psychophysics* 13:2, 247–252, URL https://doi.org/10.3758/BF03214135.

Eisner, Frank & James M McQueen (2005). The specificity of perceptual learning in speech processing. *Attention, Perception, & Psychophysics* 67:2, 224–238.

Eisner, Frank & James M McQueen (2006). Perceptual learning in speech: Stability over time. *The Journal of the Acoustical Society of America* 119:4, 1950–1953.

Jesse, Alexandra & James M. McQueen (2011). Positional effects in the lexical retuning of speech perception. *Psychonomic Bulletin & Review* 18:5, 943–950.

Kraljic, Tanya & Arthur G Samuel (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive psychology* 51:2, 141–178.

Kraljic, Tanya & Arthur G Samuel (2006). Generalization in perceptual learning for speech. *Psychonomic bulletin & review* 13:2, 262–268.

McGurk, Harry & John MacDonald (1976). Hearing lips and seeing voices. *Nature* 264:5588, 746–748.

McQueen, James M., Anne Cutler & Dennis Norris (2006). Phonological abstraction in the mental lexicon. *Cognitive Science* 30, 1113–1126.

Miller, George A. & Patricia E. Nicely (1955). An analysis of perceptual confusions among some english consonants. *The Journal of the Acoustical Society of America* 27:2, 338–352, URL http://dx.doi.org/10.1121/1.1907526.

Norris, Dennis, James M. McQueen & Anne Cutler (2003). Perceptual learning in speech. *Cognitive Psychology* 30:2, 1113–1126.

Peirce, Jonathan W (2007). Psychopy–psychophysics software in python. *Journal of neuroscience methods* 162:1, 8–13.

R Development Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org. ISBN 3-900051-07-0.

Van Linden, Sabine & Jean Vroomen (2007). Recalibration of phonetic categories by lipread speech versus lexical information. *Journal of Experimental Psychology: Human Perception and Performance* 33:6, p. 1483.